nexialog CONSULTING

Génération synthétique de données confidentielles

Comparaison de 3 algorithmes sur différents cas d'usage en banque et assurance

Henri Chhoa, Salim Kabiri et Yann Huquet
Nexialog Consulting, Paris, France

22 mai 2023

Résumé

Bien que les données issues du secteur bancaire et assurantiel soient sensibles et protégées, leur exploitation joue un rôle primordial dans la conception et l'optimisation des modèles statistiques. Les techniques de génération synthétique de données ont pour objectif de concilier ces deux aspects vis à vis de trois critères : la confidentialité, la fidélité et l'utilité. Dans cet article, nous comparons les performances de trois méthodes de synthèse de données (*Conditional Tabular Generative Adversarial Network*, *Tabular Variational Autoencoder* et copule gaussienne) sur des cas d'usage métier issus du secteur de la banque (scoring d'octroi, prédiction de revenus) et de l'assurance (tarification en prime pure). La validation de ces critères est évaluée à l'aide de scores construits à partir d'indicateurs statistiques (corrélations, distances entre distributions). Finalement, nos expérimentations ont conduit à des résultats dépendant de la nature du cas d'usage, ainsi que du type de variables présent dans les données à synthétiser.

Introduction

Le caractère confidentiel des données issues du secteur bancaire et assurantiel constitue un frein dans leur partage et leur utilisation. Cependant, dans certains cas, il est indispensable de les confier à des entités externes pour des besoins d'analyse et de modélisation.

La génération de données synthétiques à partir de données sensibles réelles représente une méthode robuste et efficace qui permet d'exploiter les informations pertinentes présentes dans une base de données en respectant la confidentialité exigée par ses propriétaires. Toutefois, si la base synthétisée est trop similaire à la base réelle, l'information qui en découle peut également toujours contenir des informations sensibles.

Les différents acteurs du marché ont élaboré de nombreuses solutions en réponse à cette problématique. Ces solutions incluent des outils de synthèse de données dotés de fonctionnalités multiples et spécifiques, qui ont été répertoriées par Nexialog Consulting lors d'une étude sur les synthétiseurs de données décrite dans [1]. Étant donné la pluralité de ces solutions, il est nécessaire de mettre en place une méthodologie d'évaluation qui permet de déterminer la méthode de synthétisation optimale pour un cas d'usage métier spécifique. [2] et [3] illustrent notamment l'application de ces méthodes pour la modélisation de risque de crédit et du défaut de carte de crédit. Par ailleurs, il existe d'autres problématiques auxquelles la synthétisation de données peut apporter une réponse, notamment la notion de déséquilibre des classes, qui est un sujet majeur dans la détection de fraude comme illustré dans [4]. En effet, dans ce domaine, les données de fraudeurs sont généralement très rares par rapport aux données non frauduleuses. Ainsi, [5] décrit l'utilisation de techniques de synthèse de données afin de générer des données synthétiques supplémentaires pour les classes minoritaires, améliorant ainsi la qualité des modèles de détection de fraude . Il est important de noter que ces sujets ne seront pas abordés dans le cadre de cet article.

L'objectif de cet article est de proposer une revue comparative des méthodes de synthèse proposées dans le package Python *Synthetic Data Vault (SDV)* [6]. Dans cette optique, nous aborderons trois exemples d'application : la notation de crédit, la prédiction du niveau de revenu et la tarification en prime pur.

Dans un premier temps, les différentes méthodes de synthétisation de données seront abordées en détail dans la Section 1. Dans un second temps, nous présenterons les critères permettant d'estimer la fidélité, la confidentialité, et l'exploitabilité des données synthétiques face aux données réelles dans la Section 2. Enfin, nous aborderons chacun des cas d'usage à travers la description des bases de données associées, les métriques employées, et l'interprétation des résultats comparatifs dans la Section 3.

1 Méthodes de synthèse de données

Nous présentons dans cette section trois différentes méthodes de synthèse traitées dans cette étude. Nous notons par la suite $\mathcal{R}$ l'ensemble des données confidentielles à synthétiser.

1.1 Copule gaussienne

Une copule à d dimensions est une fonction de répartition définie sur $[0,1]^d$ dont les lois marginales sont uniformes sur $[0,1]$ et qui permet, notamment, de simuler des données issues d'une distribution multivariée. En effet, le théorème de Sklar affirme que toute distribution jointe multivariée peut être décrite à partir des distributions marginales univariées et d'une copule qui décrit la structure de dépendance entre les variables.

Ainsi, à partir de données $X = \{\mathbf{x}^{(i)}\}_{i=1}^{N} \in \mathcal{R}$ avec $\mathbf{x} = (X_1, ..., X_P)$ vecteur aléatoire de distributions marginales $F_1, ..., F_P$, il est possible de synthétiser de nouvelles données à l'aide de la copule gaussienne et des lois marginales [1] comme il est schématisé sur la Figure 1. Le détail du processus de synthèse est décrit plus précisément dans [6].

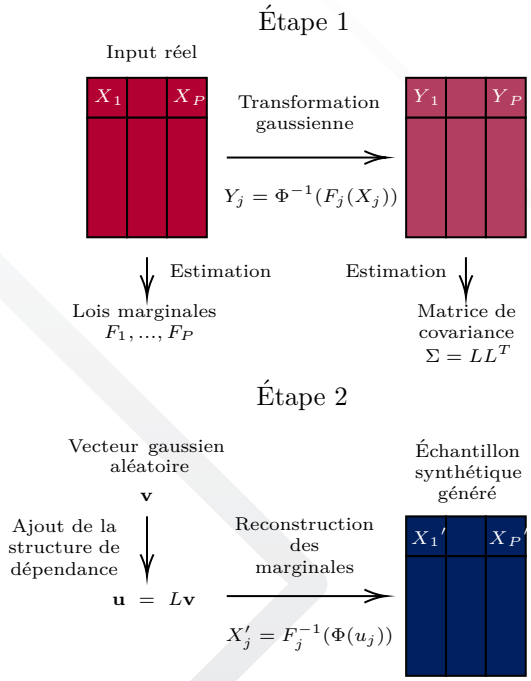

FIGURE 1 – Synthèse de données par la copule gaussienne. Φ désigne la fonction de répartition de la loi gaussienne standard et L représente la matrice issue de la factorisation de Cholesky.

1. Une transformation est appliquée sur les variables catégorielles afin de les rendre continues.

1.2 Generative Adversarial Networks (GAN)

Les Réseaux Adversariaux Génératifs ou *Generative Adversarial Networks (GAN)*, introduits pour la première fois en 2014 par une équipe de l'Université de Montréal [7], sont des algorithmes d'apprentissage automatiques permettant de générer des données synthétiques à partir d'un jeu de données réel. Contrairement aux architectures de réseaux plus classiques destinées à résoudre des problèmes de classification ou de régression, un *GAN* se distingue par l'association de deux blocs de réseaux de neurones, un générateur et un discriminateur, comme décrit dans la Figure 2.

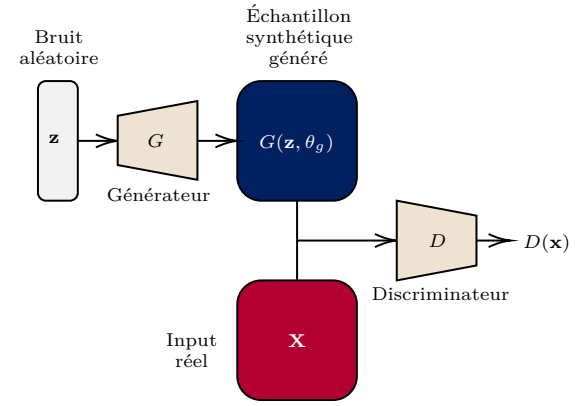

FIGURE 2 – Fonctionnement d'un *GAN*

Concrètement, l'objectif est, pour le générateur, de tromper le discriminateur en générant des données de plus en plus proches des données réelles, tandis que le discriminateur s'entraîne à mieux distinguer les données synthétiques des données réelles. Ce problème est équivalent à un problème d'optimisation de la fonction valeur $V(G, D)$ explicitée dans [7].

À partir d'un vecteur de bruit aléatoire $\mathbf{z} \sim p_{\mathbf{z}}$, le réseau de neurones correspondant au générateur G construit un échantillon synthétique $G(\mathbf{z}; \theta_g)$, où θ_g représente l'ensemble des paramètres de G. L'échantillon généré suit une distribution p_g que l'on souhaite faire approcher de la distribution réelle p_{data}. Par ailleurs, le discriminateur D génère en sortie une probabilité $D(\mathbf{x})$, qui représente la probabilité que la donnée $\mathbf{x}$ provienne des données réelles et non de la distribution générée p_g. Ainsi, D est entraîné afin de maximiser la probabilité d'assigner le bon label d'appartenance aux données d'entraînement et aux données générées par G.

L'optimisation de $V(D, G)$ est souvent effectuée récursivement par une méthode de descente de gradient stochastique divisée en 2 parties. La première partie consiste à maximiser le pouvoir discriminatoire de D en optimisant ses paramètres θ_d tout en fixant θ_g. Ensuite, les paramètres θ_g sont optimisés à leur tour par rapport à la nouvelle valeur de θ_d afin de

minimiser la détection des échantillons générés par G.

Les *GAN* étant généralement utilisés pour la génération d'images, nous privilégions dans cette étude une version dérivée appelée *CTGAN (Conditional Tabular GAN)* [8]. Cette approche permet de pallier aux contraintes imposées par les données tabulaires telles que les données mixtes, les distributions non gaussiennes, ou encore le déséquilibre des variables catégorielles, qui ne sont habituellement pas présentes dans le cas de données d'imagerie.

1.3 Variational Autoencoder (VAE)

Introduit en 2013 par *D. P. Kingma* [9], l'auto-encodeur variationnel ou *Variational Autoencoder (VAE)* est une architecture de réseaux de neurones constituée d'un encodeur, qui compresse la donnée d'entrée en une distribution de probabilité, et d'un décodeur, qui reconstruit la donnée encodée. Contrairement à un auto-encodeur traditionnel qui encode la donnée d'entrée en un vecteur fixe, un *VAE* fonctionne à l'aide d'un espace latent continu et structuré. Ceci permet la génération de nouvelles données aléatoires à partir de la distribution de probabilité estimée lors de l'apprentissage du modèle.

À partir d'un dataset réel $X = \{\mathbf{x}^{(i)}\}_{i=1}^{N} \in \mathcal{R}$ composé de N observations i.i.d. d'un vecteur aléatoire $\mathbf{x}$ caractérisé par une distribution de probabilité inconnue $\mathcal{P}(\mathbf{x})$, nous souhaitons approximer cette dernière par une distribution paramétrée p_θ de paramètres θ. On suppose que les données sont générées par un processus aléatoire impliquant une variable aléatoire continue non observée $\mathbf{z}$, qui représente un encodage latent de $\mathbf{x}$.

Très souvent, le calcul de $p_\theta(\mathbf{x})$ est coûteux ou même impossible. Il est donc nécessaire d'introduire une autre fonction q_ϕ de paramètres ϕ pour approximer la distribution a posteriori :

$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x}) \tag{1}$$

Ainsi, le problème se résume à déterminer un bon auto-encodeur probabiliste où la distribution a posteriori approximée $q_\phi(\mathbf{z}|\mathbf{x})$ est calculée par un encodeur probabiliste tandis que la distribution de vraisemblance conditionnelle $p_\theta(\mathbf{x}|\mathbf{z})$ est calculée par un décodeur probabiliste. L'entraînement du réseaux de neurones s'effectue par minimisation d'une fonction de perte, qui est équivalente à la maximisation de la fonction objectif *ELBO (Evidence Lower Bound)* explicitée dans [9]. Ce problème d'optimisation revient à minimiser la divergence entre la distribution a posteriori exact et approximée, et à maximiser la vraisemblance des données observées.

La structure d'un *VAE* dans le cas d'un encodeur gaussien est résumée dans la Figure 3.

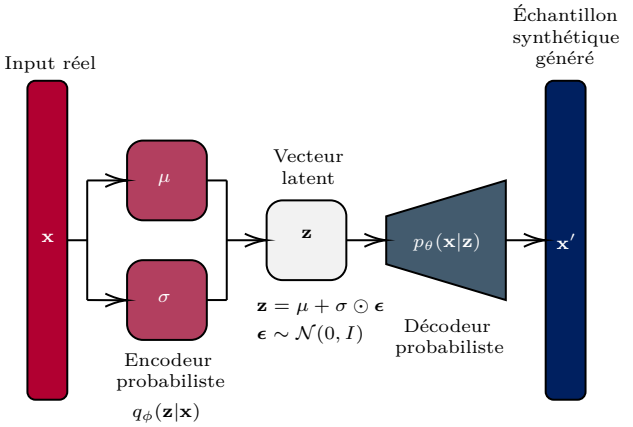

FIGURE 3 – Structure d'un VAE

Tout comme la méthode *CTGAN*, nous étudions ici une variante du *VAE* dénommée *TVAE* [8], dont l'usage a été adapté aux données tabulaires.

2 Méthodes de comparaison

Nous abordons dans cette section les métriques utilisées dans l'évaluation des performances de synthèse ainsi que les différents scores constitués.

2.1 Confidentialité des données synthétisées

Synthétiser des données permet de protéger contre les risques d'atteinte aux données réelles confidentielles. Notamment, une synthèse de données est efficace en termes de confidentialité s'il n'est pas possible de distinguer une observation réelle d'une observation synthétique, ou s'il n'est pas possible de retrouver l'information confidentielle à partir des données générées. Ainsi, il est nécessaire d'établir des mesures permettant de quantifier ce niveau de confidentialité.

Étude des corrélations inter-tables

L'une des méthodes fréquemment employées consiste à s'assurer que les corrélations des données synthétiques avec les données réelles soient minimales, comme il a été fait dans [10]. En effet, une décorrélation inter-table garantit qu'aucune inférence des données protégées ne peut être effectuée à partir des bases synthétiques.

Ces corrélations inter-tables peuvent être évaluées par le **coefficient de corrélation de Spearman** ρ_S pour les variables numériques, et par le ***V* de Cramer** pour quantifier le degré de dépendance dans le cas des variables catégorielles comme il a été fait dans [11].

Distance to Closest Record

Une autre approche de mesure de confidentialité mis en avant par [12] consiste à introduire et estimer une distance entre les individus d'une table de données réelles $X \in \mathcal{R}$ et ceux de la table synthétisée $\mathcal{S}$.

Pour $\mathbf{r} \in X$ et $\mathbf{s} \in \mathcal{S}$, cette distance est notée $d(\mathbf{s}, \mathbf{r})$. En supposant que les tables contiennent P colonnes,

$$d(\mathbf{s}, \mathbf{r}) = \sum_{j=1}^{P} d_j \tag{2}$$

tel que d_j est la distance entre la valeur de la j-ième colonne pour les individus $\mathbf{r}$ et $\mathbf{s}$. L'expression de d_j dépend de la nature de la colonne j :

— pour une colonne catégorielle,

$$d_j = \begin{cases} 0, \text{ si } s_j = r_j \\ 1, \text{ si } s_j \neq r_j \end{cases} \tag{3}$$

— pour une colonne numérique

$$d_j = |s_j - r_j| \tag{4}$$

Ainsi, la *DCR* (*Distance to Closest Record*) est définie, pour chaque individu $\mathbf{s} \in \mathcal{S}$, comme étant sa distance minimale avec tous les individus $\mathbf{r} \in X$:

$$DCR(\mathbf{s}) = \min_{\mathbf{r} \in X} d(\mathbf{s}, \mathbf{r}) \tag{5}$$

Par exemple, si s est une copie identique d'un individu r de la base d'origine, $DCR(\mathbf{s}) = 0$.

Cette distance est notamment utilisée dans la méthode empirique *holdout* décrite dans [13] et illustrée sur la Figure 4.

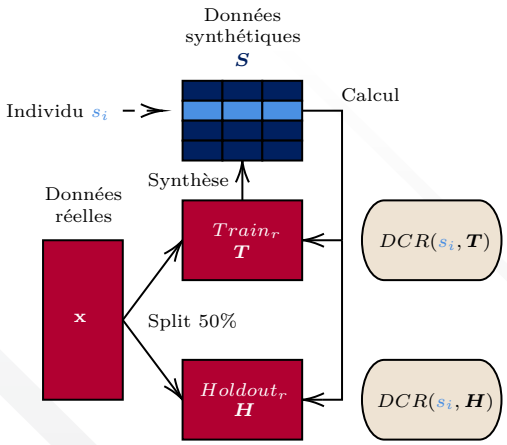

FIGURE 4 – Évaluation de la confidentialité par la *DCR*

Cette approche consiste dans un premier temps à diviser aléatoirement la base de données réelle en 2 bases de même taille, une base d'entraînement servant à la synthèse de données, et une base d'évaluation dite *holdout*. Dans un second temps, les *DCR* de chaque observation de la table synthétisée sont calculées par rapport à la base d'entraînement et la base d'évaluation. La proportion η_T des observations plus proches de la base d'entraînement que de la base d'évaluation peut ainsi être calculée. Le cas idéal, $\eta_T = 0.5$, met en évidence l'interchangeabilité des bases d'entraînement et de *holdout* vis à vis des données synthétisées. Par conséquent, il ne serait pas possible de déterminer l'origine d'une observation synthétisée ce qui assurerait le critère de confidentialité.

Méthode du score de propension

Snoke et al. (2018) présentent dans [14] la méthode du score de propension, permettant de quantifier la capacité d'un modèle à distinguer une observation synthétique d'une observation réelle. Cette méthode consiste à créer une nouvelle base de données en mélangeant les données réelles et synthétiques, chacune de taille N, formant ainsi une base de taille $2N$. Un label d'appartenance est attribué à chaque observation selon leur origine (0 pour les données réelles et 1 pour les données synthétiques). Ensuite, la probabilité d'appartenir aux données synthétiques $\hat{p}_i$ estimée par un classifieur permet de construire la *propensity Mean Square Error (pMSE)* :

$$pMSE = \frac{1}{2N} \times \sum_{i=1}^{2N} (\hat{p}_i - c)^2 \tag{6}$$

où c représente la proportion de données synthétiques, ici $c = 0.5$. Le cas idéal correspond au cas où toutes les prédictions $\hat{p}_i$ sont égales à c, indiquant ainsi que les données synthétiques et réelles sont complètement indistinguables par le classifieur.

Score global de confidentialité

À ce stade, nous définissons un score global de confidentialité qui prend en compte les 3 métriques définies précédemment, afin de faciliter la comparaison entre les modèles. Ce **Privacy Score** est calculé de la manière suivante :

$$\textbf{Privacy Score} = \frac{\text{Score 1} + \text{Score 2} + \text{Score 3}}{3} \tag{7}$$

Le *Score 1* est une agrégation des métriques de corrélation inter-tables introduites précédemment. L'idée est de construire une valeur comprise entre 0 et 100 % qui soit d'autant plus élevée que les corrélations inter-tables sont faibles.

Pour cela, nous appliquons la transformation $x \in [0, 1] \mapsto 1 - x \in [0, 1]$ à la moyenne des corrélations des variables numériques ρ_S et catégorielles V.

$$\text{Score 1} = 1 - \frac{1}{2} \cdot \left[\frac{\sum_{j \in \mathcal{N}} |\rho_S(R_j, S_j)|}{p_{num}} + \frac{\sum_{j \in \mathcal{C}} V(R_j, S_j)}{p_{cat}} \right] \tag{8}$$

avec $(R_j)_{1 \leq i \leq P}$ et $(S_j)_{1 \leq i \leq P}$ respectivement les colonnes des tables réelles et synthétiques, $\mathcal{N}$ l'ensemble des indices des colonnes numériques (p_{num} indices), $\mathcal{C}$ l'ensemble des indices des colonnes catégorielles (p_{cat} indices) [1] et V le V de Cramer.

1. $\mathcal{N} \cup \mathcal{C} = [\![1\ ;\ P]\!]$, avec P le nombre de colonnes.

Le *Score 2* résulte de η_T, la part des observations synthétisées qui sont plus proches de la base d'entraînement que de la base *holdout*. Comme mentionné précédemment, un niveau de confidentialité idéal est atteint lorsque $\eta_T = 0.5$. Ainsi, le *Score 2* est calculé de la façon suivante :

$$\text{Score 2} = \begin{cases} 2 \cdot \eta_T \text{ , si } \eta_T \leq 0.5 \\ 2 - 2 \cdot \eta_T \text{ , sinon} \end{cases} \tag{9}$$

Enfin, le *Score 3* est construit à partir de la *pMSE* de telle sorte à avoir un score compris entre 0 et 100%, et maximal lorsque *pMSE=0*.

$$\text{Score 3} = 1 - 4 \cdot pMSE \tag{10}$$

2.2 Fidélité des données synthétisées

Au même titre que la confidentialité, la fidélité des données est un critère primordial. Il est important de noter que la fidélité et la confidentialité des données peuvent être discordants, puisqu'augmenter la confidentialité d'un jeu de données synthétique a tendance à diminuer sa fidélité statistique.

L'évaluation de la fidélité des données synthétiques est donc également un processus important pour garantir la pertinence des échantillons générés par les synthétiseurs. Elle inclut notamment la comparaison des corrélations intra-tables et l'analyse des distributions.

Étude des corrélations intra-tables

Si dans la partie 2.1, la corrélation entre les colonnes des tables synthétisées et les tables réelles devait être minimisée afin de maximiser le niveau de confidentialité, ici, les corrélations intra-tables (évaluées pour chaque couple de colonnes au sein de la table synthétisée et au sein de la table réelle) doivent être similaires dans le but de préserver la structure de dépendance des variables [15]. Pour cela, les **matrices de corrélation** [1] des tables synthétisées et réelles sont calculées.

Similarité des distributions

Afin de quantifier la similarité des distributions des variables originales $(R_j)_{1\leq i\leq p}$ par rapport aux variables synthétiques $(S_j)_{1\leq i\leq p}$, nous procédons à des tests statistiques selon le type de variable (catégorielle ou numérique).

Dans le cas des variables catégorielles, le score **TV-Complement** (**TVC**) décrit dans [16] et construit à partir de la distance en variation totale (*Total Variation Distance*), est utilisé. Pour chaque couple (R_j, S_j), les densités de probabilité de toutes les modalités possibles $\omega \in \Omega_j$ d'une colonne j sont calculées. Le score final s'exprime comme suit :

$$\begin{aligned} TVC(R_j, S_j) &= 1 - \delta(R_j, S_j) \\ &= 1 - \frac{1}{2} \sum_{\omega \in \Omega_i} |\mathbb{P}(R_j = \omega) - \mathbb{P}(S_j = \omega)| \end{aligned} \tag{11}$$

Pour les variables numériques, la statistique de *Kolmogorov-Smirnov* est calculée. Elle évalue l'écart maximal entre les fonctions de répartition des variables $(R_j)_{1\leq j\leq P}$ et $(S_j)_{1\leq j\leq P}$. Le score constitué est appelé **KSComplement** (**KSC**) dont l'expression est définie comme suit :

$$\begin{aligned} KSC(R_j, S_j) = 1 - \max_{\omega \in \Omega_j} &|\mathbb{P}(R_j \leq \omega) \\ &- \mathbb{P}(S_j \leq \omega)| \end{aligned} \tag{12}$$

Score global de fidélité

À l'instar du score global de confidentialité, nous introduisons le score global de fidélité (**Fidelity Score**) calculé à partir des distances de corrélations intra-tables ainsi que les similarités des distributions :

$$\textbf{Fidelity Score} = \frac{\text{Score 4} + \text{Score 5}}{2} \tag{13}$$

Le *Score 4* est très similaire au *Score 1* dans le sens où nous agrégeons des corrélations tout en contraignant sa valeur entre 0 et 100%. La seule différence ici est que nous introduisons des distances de corrélations (entre les tables synthétisées et les tables réelles). Ces distances sont calculées à partir des matrices de corrélations intra-tables (corrélation de Spearman pour les variables numériques et les V de Cramer pour les variables catégorielles). La somme des distances des corrélations dans le cas des variables numériques est normalisée par $\frac{1}{2\times(P_{num})^2}$ car chacune de ces $(P_{num})^2$ distances est comprise entre 0 et 2. Les distances des corrélations catégorielles, quant à elles, sont uniquement normalisée par $\frac{1}{(P_{cat})^2}$ car elles sont toutes comprises entre 0 et 1.

$$\begin{aligned} \text{Score 4} = 1 - \frac{1}{2}\Bigg[&\frac{1}{2 \times (P_{num})^2} \cdot \left(\sum_{(i,j)\in\mathcal{N}^2} |\rho_S(S_i, S_j) - \rho_S(R_i, R_j)| \right) \\ &+ \frac{1}{(P_{cat})^2} \cdot \left(\sum_{(k,l)\in\mathcal{C}^2} |V(S_k, S_l) - V(R_k, R_l)| \right) \Bigg] \end{aligned} \tag{14}$$

avec $\mathcal{N}$ l'ensemble des indices des colonnes numériques (p_{num} indices), $\mathcal{C}$ l'ensemble des indices des colonnes catégorielles (p_{cat} indices) et V le V de Cramer.

1. Matrices de V de Cramer pour les variables catégorielles.

Pour calculer le *Score 5*, nous regroupons la contribution des *TVC* et *KSC* en moyennant ces deux indicateurs, produisant ainsi un score compris entre 0 et 100%.

$$\text{Score } 5 = \frac{1}{2} \cdot \left[\frac{\sum_{j \in \mathcal{N}} TVC(R_j, S_j)}{p_{num}} + \frac{\sum_{j \in \mathcal{C}} KSC(R_j, S_j)}{p_{cat}} \right] \tag{15}$$

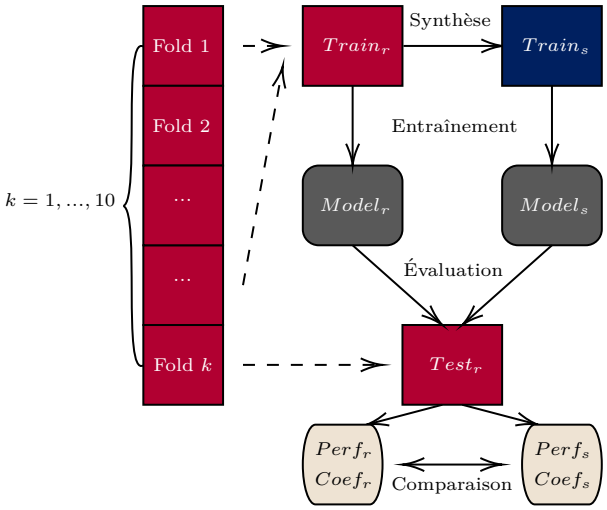

FIGURE 5 – Validation croisée 10-*fold*

2.3 Stratégie d'évaluation des performances sur un cas d'usage métier

En plus des critères de fidélité et de confidentialité énoncés, il est nécessaire de s'assurer que les données synthétiques aient un comportement similaire aux données réelles lors de leur utilisation dans un cas d'usage spécifique. Nous définissons dans ce qui suit les principales méthodes d'évaluation utilisées dans le cas d'une régression et d'une classification.

2.3.1 Validation croisée

La validation croisée *k-fold* est une méthode d'évaluation de la qualité du processus de modélisation. Elle consiste à diviser le jeu de données initial en k paquets distincts, et à réaliser l'ensemble du processus de modélisation séquentiellement. À chaque itération, on considère un paquet différent pour l'échantillon de test, le reste servant à l'échantillon d'apprentissage. Ainsi, les k itérations de la procédure résultent en k résultats de performance, qui peuvent être agrégés par la suite pour analyse. Cette méthode permet d'évaluer la capacité de généralisation d'un modèle sur des données *out of sample* tout en s'assurant que chaque observation appartienne aux bases d'apprentissage et d'évaluation. Par conséquent, les performances du modèle peuvent être estimées de manière plus robuste et non biaisée, contrairement à une méthodologie basée sur un seul découpage aléatoire en deux bases *train*/*test*.

Nous adaptons cette méthodologie à notre étude en incorporant une étape de synthétisation de données, de manière à pouvoir comparer sur un cas d'usage les performances du modèle entraîné sur les données réelles et synthétiques. La procédure de validation croisée 10-*fold* mise en place dans cette étude est détaillée en Figure 5.

2.3.2 Utilité des données synthétisées

L'exploitabilité des données synthétiques est grande si elles permettent d'obtenir des résultats de performance assez similaires dans les deux cas, ce qui traduirait une proximité de l'information présente dans la base réelle et la base synthétique. De la même façon, l'importance des variables du modèle final se doit d'être assez similaire à celle du modèle entraîné sur les données réelles. En l'occurrence, dans le cas d'une régression linéaire ou logistique que nous considérons dans cette étude, il est possible de moyenner les estimations de coefficients et d'évaluer leurs écarts-types à l'issue de la procédure de validation croisée afin de comparer leurs valeurs en fonction de la méthode de synthèse utilisée.

Le modèle entraîné sur le jeu de données synthétique doit donc idéalement présenter les mêmes performances que le modèle entraîné sur le jeu réel ainsi que les mêmes importances de variables.

Métriques de classification

Plusieurs métriques permettent de mesurer les performances d'un modèle de Machine Learning. Dans le cadre d'une classification, les métriques du *F1-score* et de la *Receiver Operating Characteristic Area Under Curve (ROC AUC)* sont utilisées dans cette étude.

Le *F1-score* correspond à la moyenne harmonique entre la *precision* (taux de prédictions correctes parmi toutes les prédictions positives) et le *recall* (taux de prédictions correctes parmi tous les cas positifs). Cette métrique est particulièrement utilisée pour les problèmes utilisant des données déséquilibrées comme la détection de fraudes ou la prédiction d'incidents graves. Elle s'intéresse aux classes prédites, et est donc dépendante du seuil de classification fixé. Une probabilité prédite supérieure ou inférieure à 0.5 correspondant respectivement à une prédiction de la

classe positive ou négative.

Par opposition, la *ROC AUC* est une métrique non dépendante du seuil communément utilisé dans l'évaluation des performances de classification. Cette métrique est égale à l'aire sous la courbe *ROC*, qui représente le taux de vrais positifs en fonction du taux de faux positifs.

Un score est calculé pour chacune de ces métriques [1] afin de comparer les performances du modèle entraîné sur les données synthétiques M_s au modèle entraîné sur le modèle M_r, défini comme :

$$\text{Score } 6 = 1 - |ROCAUC_{M_r} - ROCAUC_{M_s}| \quad (16)$$

$$\text{Score } 7 = 1 - |F1\text{-}score_{M_r} - F1\text{-}score_{M_s}| \quad (17)$$

Métriques de régression

Dans le cadre d'une régression, les métriques de la *Root Mean Square Error (RMSE)* et de la *Mean Absolute Error (MAE)* sont utilisées.

La ***RMSE*** est la racine carrée de l'erreur quadratique moyenne et est définie comme :

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (18)$$

où y représente les valeurs observées et $\hat{y}$ les valeurs prédites de la variable d'intérêt.

La ***MAE***, ou erreur absolue moyenne, est la moyenne des valeurs absolues des erreurs, définie comme :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (19)$$

De même que pour la classification, un score est calculé pour chacune de ces métriques afin de comparer les performances du modèle M_s au modèle M_r.

$$\text{Score } 8 = 1 - |RMSE_{M_r} - RMSE_{M_s}| \quad (20)$$

$$\text{Score } 9 = 1 - |MAE_{M_r} - MAE_{M_s}| \quad (21)$$

Estimation des coefficients d'importance des variables dans chaque modèle de synthèse

Les variables explicatives utilisées dans les cas d'usage (classification ou régression) ont souvent des importances différentes. La méthode de calcul de ces importances dépend du modèle utilisé dans la prédiction des variable cibles. Dans notre cas, nous avons fait le choix d'utiliser des modèles de régression (régression logistique pour les problèmes de classifications et modèles linéraires généralisés « *Generalized Linear Models - GLM* » pour l'estimation de variables continues). Ainsi, la différence entre l'estimation des coefficients $\beta_{s,1}, ..., \beta_{s,P}$ issus du modèle M_s et les coefficients $\beta_{r,1}, ..., \beta_{r,P}$ du modèle M_r est calculée comme suit :

$$d = \frac{1}{P} \cdot \sum_{i=1}^{P} \frac{|\beta_{s,i} - \beta_{r,i}|}{\Delta_{max}} \quad (22)$$

où Δ_{max} est la valeur maximale des $|\beta_{s,i} - \beta_{r,i}|$ observée pour les trois modèles de synthèse considérés.

Par la suite, nous constituons le score d'importance qui n'est autre qu'une transformée de l'expression précédente afin d'obtenir un score croissant inclus dans $[0, 1]$:

$$\text{Score } 10 = 1 - d \quad (23)$$

Score global d'utilité

Tout comme les scores de confidentialité et de fidélité, nous introduisons le score global d'utilité (**Utility Score**) calculé à partir des indicateurs précédemment évoqués :

— cas d'une classification :

$$\textbf{Utility Score} = \frac{\text{Score } 6 + \text{Score } 7 + \text{Score } 10}{3} \quad (24)$$

— cas d'une régression :

$$\textbf{Utility Score} = \frac{\text{Score } 8 + \text{Score } 9 + \text{Score } 10}{3} \quad (25)$$

3 Application aux cas d'usage

Dans cette partie, nous présentons l'application des méthodes de synthèse sur trois cas d'usage métier différents : le scoring en octroi de crédit, la prédiction du niveau de revenu et la modélisation du coût pour la tarification de la prime pure en assurance. Nous analysons ensuite les résultats comparatifs obtenus à partir des différents scores et interprétons ces résultats.

3.1 Scoring en octroi de crédit

L'octroi de crédit nécessite une anticipation des risques auxquels s'expose la banque face à un emprunteur pouvant faire défaut sur son contrat de prêt. La mise en place d'un scoring permet de créer des classes de risque et de quantifier cette probabilité de défaut à partir des informations personnelles des emprunteurs.

1. Les *Score 6* et *Score 7* sont dans $[0, 1]$ car $0 \leq ROCAUC \leq 1$ et $0 \leq F1\text{-}score \leq 1$.

3.1.1 Métrique du cas d'usage et données

La base de données s'intitulant *Credit Risk Dataset* considérée dans cette sous-section provient initialement du site *Kaggle*, et a été traitée de manière à discrétiser l'ensemble des variables comme décrit dans [17] afin de construire une grille de notation selon les modalités des variables. Finalement, la base obtenue est composée de 32581 observations et 18 variables explicatives discrètes. Nous avons alors modélisé la probabilité de défaut par un modèle de régression logistique et évalué ses performances à l'aide des métriques de classification décrites précédemment.

3.1.2 Résultats et analyses

Les résultats de la validation croisée pour le scoring en octroi de crédit sont détaillés dans la Figure 6.

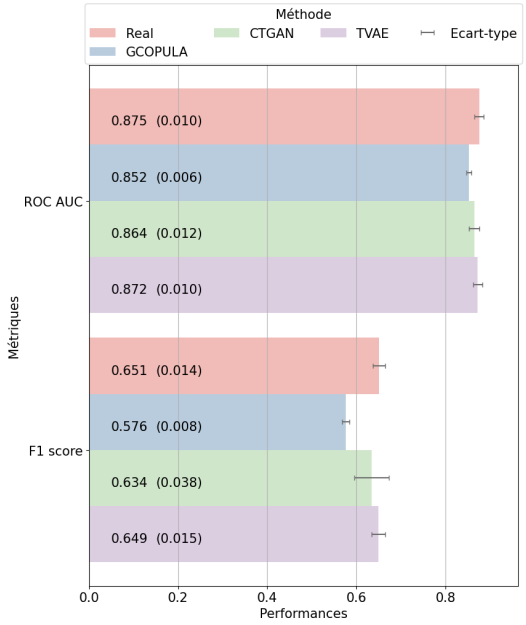

FIGURE 6 – Résultats agrégés de la validation croisée 10-*fold* pour les métriques de *ROC AUC* et *F1-score* appliquée au cas du *credit scoring*. Pour chaque modèle de synthèse, les estimations moyennes et les écarts-types entre parenthèses sont données. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Pour les deux métriques considérées, les performances du modèle entraîné sur les données synthétisées par le *TVAE* sont les plus proches des performances du modèle entraîné sur données réelles avec un écart-type comparable. D'un autre côté, la méthode de synthèse *CTGAN* permet d'obtenir des résultats moyens très similaires, avec en revanche une variabilité plus grande comme l'indique les valeurs d'écart-type. Ainsi, cette faible stabilité témoigne d'un surapprentissage du modèle qui est très sensible aux données d'entraînement.

Par ailleurs, la méthode de copule gaussienne conduit à des performances très en dessous des performances issues de données réelles, ce qui tend à montrer l'inadéquation de cette méthode dans cette application. Cela peut être expliqué par le fait que la totalité des variables sont catégorielles dans ce cas de scoring. En effet, comme expliqué dans la section 1.1, les données catégorielles sont transformées en données continues à l'aide d'un échantillonnage sur plusieurs lois gaussiennes tronquées. Ainsi, d'une part, cela pourrait dégrader l'estimation des lois marginales associées car les variables ne suivent plus leurs distributions originales. D'autre part, l'estimation des lois marginales s'effectue à l'aide de lois usuelles comme indiqué dans la section 1.1. Or, la transformation des distributions empiriques des variables catégorielles ne permet pas d'obtenir *in fine* des distributions usuelles.

Les coefficients de la régression logistique sont quant à eux assez différents selon la méthode de synthèse employée comme le montre la Figure 7.

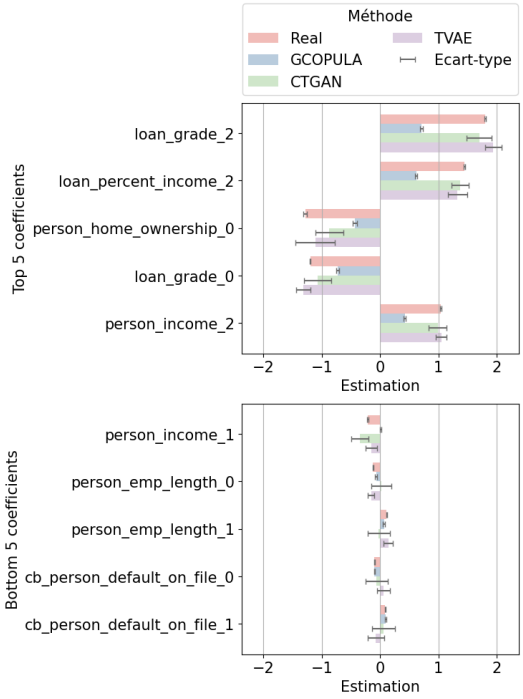

FIGURE 7 – Estimation moyenne des 5 plus/moins grands coefficients issus de la validation croisée 10-*fold* avec les écarts-types associés. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Globalement, le *TVAE* permet d'obtenir les estimations les plus proches du cas réel avec un écart-type raisonnable. La copule gaussienne, au contraire, aboutit à des coefficients très sous-estimés, bien qu'ayant des écarts-types proches du cas réel. Enfin, les coefficients obtenus avec le *CTGAN* se rapprochent également des coefficients du cas réel, mais les écarts-types sont plus élevés que dans le cas du *TVAE*. Il est intéressant de remarquer que les très faibles

coefficients sont mal estimés dans le cas du *CTGAN* et du *TVAE*, ce qui indique une difficulté à capturer la faible importance de ces variables pour ces méthodes de synthèse.

Le récapitulatif des scores de qualité est donné dans la Table 1, ci-dessous.

Scores	Gaussian Copula	CTGAN	TVAE
Corrélations inter-tables (Score 1)	100 %	100 %	99.99 %
DCR (Score 2)	2.95 %	1.19 %	0.60 %
pMSE (Score 3)	92.74 %	98.20 %	99.70 %
Privacy Score	**65.23** %	**66.46** %	**66.76** %
Corrélations intra-tables (Score 4)	99.15 %	99.26 %	99.77 %
TVC/KSC (Score 5)	91.90 %	95.69 %	98.86 %
Fidelity score	**95.53** %	**97.47** %	**99.32** %
ROC AUC adapté (Score 6)	97.90 %	98.90 %	99.70 %
F1-score adapté (Score 7)	92.50 %	98.30 %	99.80 %
Feature importance (Score 10)	63.13 %	87.72 %	92.39 %
Utility Score	**84.44** %	**94.97** %	**97.3** %

TABLE 1 – Résumé des scores des différentes méthodes de synthèse pour le cas du scoring en octroi de crédit.

Nous pouvons remarquer que les scores de la *DCR* sont tous très faibles, quelque soit la méthode de synthèse utilisée, ce qui traduit une très forte proximité entre les données synthétiques et les données réelles. En effet, la base de données étudiée ici est uniquement composée de variables catégorielles, donc ces dernières prennent un nombre fini de valeurs. Cela mène inévitablement à la réplication exacte d'observations dans les données synthétisées, réduisant ainsi drastiquement le score lié à la *DCR*.

De plus, les trois méthodes de synthèse permettent d'obtenir globalement de très bons scores de fidélité statistique. Enfin, les score issus de la *ROC AUC* et du *F1-score* sont également très bons malgré une légère sous-performance de la copule gaussienne. Enfin, le score d'importance de variables démontrent une autre faiblesse de cette dernière, avec un score très dégradé comparé aux deux autres méthodes, ce qui est conforme aux observations faites sur la Figure 7.

3.2 Prédiction du niveau de revenu

Le niveau de revenu peut être une information importante dans la connaissance de son portefeuille et la détermination de la valeur client. Ainsi, il peut être intéressant de modéliser cette variable d'intérêt à partir des caractéristiques démographiques des prospects, tels que l'âge, le métier, le statut marital, ou encore le niveau d'éducation.

3.2.1 Métrique du cas d'usage et données

La base de données étudiée ici et intitulée *Adult income dataset* provenant du site *Kaggle* met en lumière ce cas d'usage précis. À partir d'une base composée de 30139 observations, 53 variables explicatives continues et catégorielles, nous nous intéressons à la prédiction de la variable d'intérêt binaire dont la classe positive indique un revenu supérieur à 50000 $. Dans ce cas de classification, nous avons modélisé le niveau de revenu par un modèle de régression logistique et évalué ses performances par la *ROC AUC* et le *F1-score*. À noter que les observations dupliquées ou contenant des valeurs manquantes ont été exclues de l'analyse.

3.2.2 Résultats et analyses

Pour ce cas d'usage, les résultats de la validation croisée sont détaillés dans la Figure 8. Nous constatons que le modèle lié à la copule gaussienne produit les performances les plus éloignées, tandis que les méthodes *CTGAN* et *TVAE* présentent des performances comparables. Encore une fois, la présence de multiples variables catégorielles peuvent expliquer cet écart de performance.

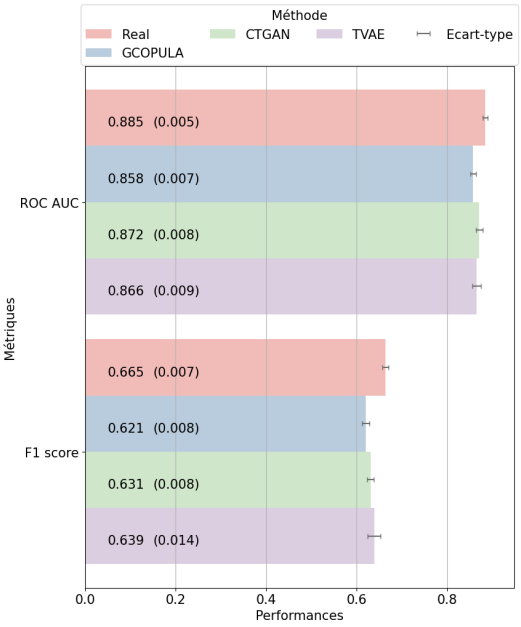

FIGURE 8 – Résultats agrégés de la validation croisée 10-*fold* pour les métriques de *ROC AUC* et *F1-score* appliquée au cas de la prédiction du niveau de revenu. Pour chaque modèle de synthèse, les estimations moyennes et les écarts-types entre parenthèses sont données. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Les estimations des cinq plus grands et plus petits coefficients sont illustrées dans la Figure 9.

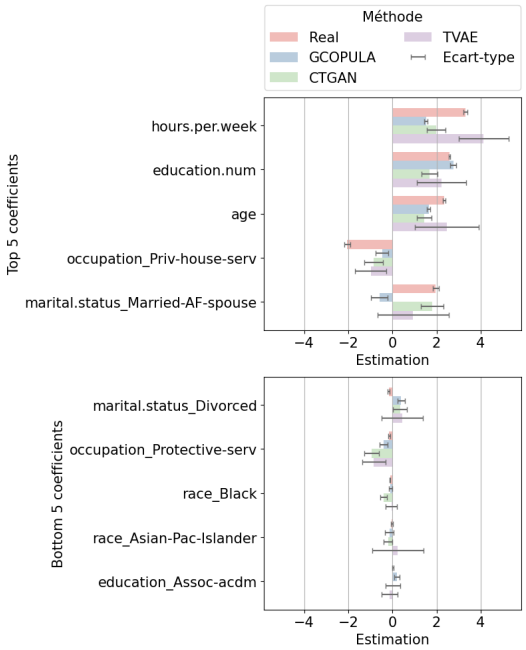

FIGURE 9 – Estimation moyenne des 5 plus/moins grands coefficients issus de la validation croisée 10-*fold* avec les écarts-types associés. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Les écarts d'estimation sont très variables selon les méthodes de synthèse et au sein d'une même méthode, ce qui ne permet pas de déterminer clairement la méthode optimale dans ce cas d'usage. Par exemple, nous pouvons remarquer des signes opposés pour la variable « divorcé » représentée par « marital.status_Divorced » dans la Figure 9, quelque soit la méthode de synthèse employée. Également, la dispersion des estimations issues du *TVAE* illustrée par les écarts-types importants indiquent une faible robustesse de cette méthode qui ne permet pas de généraliser ses estimations de manière stable.

Le récapitulatif des scores est donné dans la Table 2.

Scores	Gaussian Copula	CTGAN	TVAE
Corrélations inter-tables (Score 1)	99.93 %	99.93 %	99.93 %
DCR (Score 2)	27.85 %	30.88 %	31.32 %
pMSE (Score 3)	40.32 %	71.67 %	62.24 %
Privacy Score	**56.03** %	**67.50** %	**64.50** %
Corrélations intra-tables (Score 4)	97.34 %	98.97 %	99.01 %
TVC/KSC (Score 5)	86.51 %	90.65 %	89.27 %
Fidelity score	**91.93** %	**94.81** %	**94.14** %
ROC AUC adapté (Score 6)	97.34 %	98.69 %	98.08 %
F1-score adapté (Score 7)	95.56 %	96.61 %	97.43 %
Feature importance (Score 10)	73.91 %	85.98 %	80.11 %
Utility Score	**88.94** %	**93.76** %	**91.87** %

TABLE 2 – Résumé des scores des différentes méthodes de synthèse pour le cas de la prédiction du niveau de revenu.

La méthode *CTGAN* permet d'obtenir les meilleures performances sur les trois critères d'évaluation. Nous remarquons que les scores liés à la *DCR* sont toujours assez faibles, mais bien supérieurs au cas précédent de scoring en octroi de crédit. Ceci est expliqué par la présence d'un plus grand nombre de variables, bien qu'elles soient discrètes, diminuant ainsi la probabilité de réplication d'observations réelles dans les tables synthétisées. Dans l'ensemble, les scores obtenus par la copule gaussienne sont inférieurs aux autres méthodes, avec notamment un score issu de la *pMSE* très faible et donc une confidentialité réduite. Par ailleurs, les scores de fidélité et d'utilité sont relativement proches et élevés, avec néanmoins une dégradation des scores associés aux estimations de coefficients.

3.3 Modélisation du coût pour la tarification de la prime pure

Une étude est réalisée chez un assureur dans un contexte de prévoyance santé individuelle et vise à déterminer le cashback à allouer aux contrats Santé des Travailleurs Non Salariés (TNS). Ce dernier est défini comme une partie des primes reversée aux assurés en cas de non consommation. Afin de pouvoir déterminer un cashback cohérent et adapté à chaque client, il est nécessaire de s'intéresser à la modélisation de la prime pure. Cette prime quantifie le montant que l'assureur doit facturer à l'assuré afin de couvrir les risques de sinistre.

3.3.1 Métrique du cas d'usage et données

Plus précisément, la tarification de la prime pure est construite comme le produit entre la fréquence et le coût liés à un sinistre. La modélisation de ces deux composantes s'effectue à l'aide de modèles linéaires généralisés GLM (*Generalized Linear Models*), qui permettent de relier les variables explicatives et la variable cible via une fonction lien.
Ainsi, à partir d'une base de données composée de 16921 observations et 4 variables explicatives, nous avons modélisé le logarithme du coût moyen de la prime de l'assuré par un *GLM*. Ce modèle a ensuite été évalué à l'aide des métriques de $RMSE$ et MAE introduites précédemment.

3.3.2 Résultats et analyses

Les performances de la validation croisée pour le modèle GLM sont détaillées dans la Figure 10.

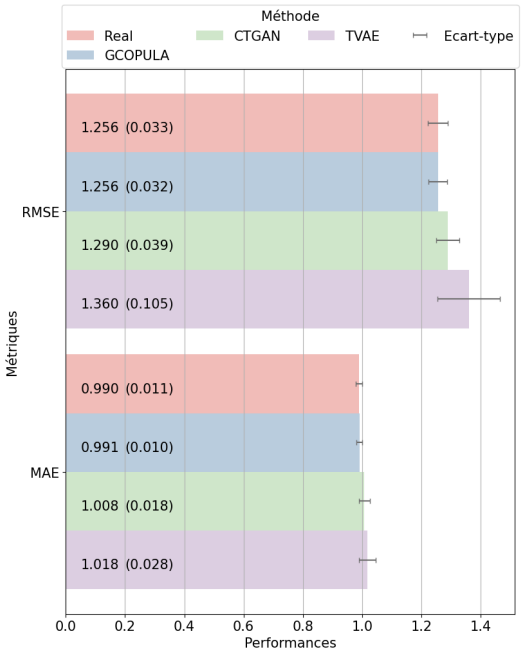

FIGURE 10 – Résultats agrégés de la validation croisée 10-*fold* pour les métriques de *RMSE* et *MAE* appliquée au cas de tarification en prime pure. Pour chaque modèle de synthèse, les estimations moyennes et les écarts-types entre parenthèses sont données. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Pour les deux métriques considérées, les performances des modèles entraînés sur les données synthétisées par la copule gaussienne et le *CTGAN* sont plus proches des performances du modèle entraîné sur données réelles que pour la méthode *TVAE*, avec en particulier des résultats très similaires pour la méthode de copule gaussienne. Nous pouvons noter que dans ce cas également, la méthode *TVAE* engendre des écarts-types importants.

La Figure 11 illustre la faible stabilité des estimations issues des méthodes *CTGAN* et *TVAE*, signe de surapprentissage de ces algorithmes. Celles-ci ont des écarts-types très largement supérieures à l'écart-type de la méthode de copule gaussienne, qui se rapproche du cas des données réelles. La robustesse de ces méthodes de synthèse est réduite dans ce cas d'usage, provoquant à chaque itération de validation croisée des données synthétiques relativement différentes, et donc des estimations de coefficients très variables. Nous pouvons également observer de forts écarts dans l'estimation moyenne des coefficients, avec notamment un signe opposé pour le coefficient lié à la variable du sexe. Ces écarts importants peuvent être expliqués par le volume de données réduit en termes d'observations et du faible nombre de variables, ce qui expliquerait la sous-performance des méthodes de *CTGAN* et *TVAE* construites à partir de réseaux de neurones profonds, qui nécessitent un volume de données important.

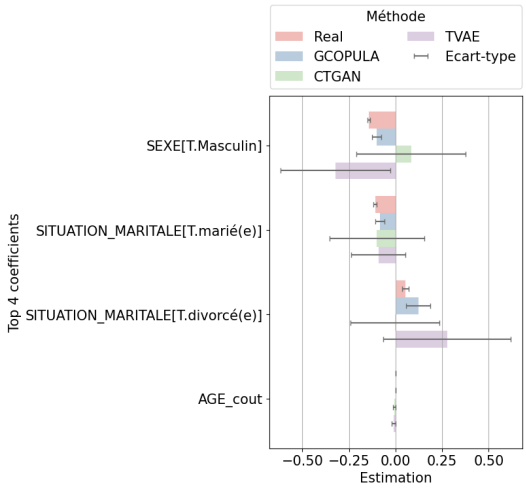

FIGURE 11 – Estimation moyenne des coefficients (hors *intercept*) issus de la validation croisée 10-*fold* avec les écarts-types associés. Ces derniers sont représentés par les barres d'erreurs dont les longueurs correspondent à 2 fois les écarts-types.

Le récapitulatif des scores de confidentialité, fidélité et utilité est donné dans la Table 3.

Scores	**Gaussian Copula**	**CTGAN**	**TVAE**
Corrélations inter-tables (Score 1)	99.69 %	99.81 %	99.34 %
DCR (Score 2)	83.74 %	85.44 %	74.85 %
pMSE (Score 3)	97.40 %	91.62 %	94.58 %
Privacy Score	**93.61** %	**92.29** %	**89.59** %
Corrélations intra-tables (Score 4)	96.43 %	96.96 %	97.08 %
TVC/KSC (Score 5)	88.45 %	83.27 %	86.37 %
Fidelity score	**92.44** %	**90.11** %	**91.72** %
RMSE adaptée (Score 8)	100 %	96.60 %	89.6 %
MAE adaptée (Score 9)	99.90 %	98.20 %	97.2 %
Feature importance (Score 10)	87.24 %	55.20 %	55.12 %
Utility Score	**95.71** %	**83.33** %	**80.64** %

TABLE 3 – Résumé des scores des différentes méthodes de synthèse pour le cas de tarification en prime pure.

Globalement, la méthode de copule gaussienne est meilleure sur les trois axes de comparaison. Nous pouvons remarquer que les scores issus de la *DCR* sont très élevés et bien supérieurs que dans les deux cas précédents. Cela s'explique par le fait que la variable du coût moyen est continue, et donc prend des valeurs dans un espace infini, évitant ainsi la réplication exacte d'observations dans les tables synthétisées.

Bien qu'ayant des performances très similaires, on retrouve la faible performance des méthodes *CTGAN* et *TVAE* sur l'estimation des coefficients, comme expliqué ci-dessus. Cela dégrade ainsi assez fortement le score global d'utilité.

4 Conclusion

Les méthodes de synthèse étudiées conduisent à des résultats assez variables selon le cas d'usage considéré,

ne permettant pas de définir une unique méthode optimale au global.

Dans l'ensemble, les trois méthodes de synthèse produisent des données fidèles aux données réelles d'un point de vue statistique, avec des distributions et corrélations proches comme l'indiquent les scores de fidélité. Outre l'estimation des coefficients de régression qui est assez instable, les trois méthodes permettent d'obtenir des résultats de performance proches pour l'ensemble des cas d'usage traités, témoignant ainsi de l'exploitabilité des données synthétiques.

En particulier, les résultats liés au critère de confidentialité sont très dépendants du cas d'usage. En effet, la présence de données catégorielles ou numériques discrètes peut potentiellement entraîner la réplication de certaines observations lors de la synthèse de données. Ainsi, indépendamment de la méthode de synthèse utilisée, la synthétisation de données admet une faiblesse sur la confidentialité des données générées dans le cas de données discrètes. Au contraire, la présence de données numériques continues permet de corriger ce défaut et d'éviter la duplication exacte de certaines observations.

Par ailleurs, la méthode de copule gaussienne semble moins adaptée lorsque les données contiennent un nombre non négligeable de variables catégorielles. Cet écart est potentiellement expliqué par la phase de prétraitement permettant de rendre continu les variables catégorielles. L'estimation des lois marginales est alors dégradée, conduisant inévitablement à un processus de synthétisation biaisé et d'autant plus accentué que les variables catégorielles sont nombreuses.

D'autre part, l'efficacité des méthodes de *CTGAN* et de *TVAE* semblent réduite lorsque le volume de données est faible. Par opposition, la méthode de copule gaussienne ne semble pas être affectée par ce manque de données. Cet écart se retrouve dans la variabilité des estimations de performances et de paramètres. En effet, les méthodes *CTGAN* et *TVAE* semblent assez peu robustes au vu des écarts-types importants malgré de bons résultats en moyenne. Ce déséquilibre biais-variance indique un sur-apprentissage des données d'entraînement.

Enfin, il serait intéressant de se pencher sur l'optimisation des modèles de synthèse afin d'ajuster la qualité des données synthétiques selon un ou plusieurs critères en particulier. Dans certains cas, une confidentialité accrue peut être requise au détriment d'une exploitabilité dégradée. Au contraire, dans d'autres cas, valoriser l'exploitabilité des données selon l'usage métier considéré est préférable, au détriment du critère de confidentialité ou de fidélité.

Références

[1] Yann Huquet. Benchmark data anonymisation. *Nexialog Consulting*, 2023.

[2] Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A Ríos, and Manuel Graña. Assessment of creditworthiness models privacy-preserving training with synthetic data. In *Hybrid Artificial Intelligent Systems : 17th International Conference, HAIS 2022, Salamanca, Spain, September 5–7, 2022, Proceedings*, pages 375–384. Springer, 2022.

[3] Belén Vega-Márquez, Cristina Rubio-Escudero, José C Riquelme, and Isabel Nepomuceno-Chamorro. Creation of synthetic data with conditional generative adversarial networks. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019) Seville, Spain, May 13–15, 2019, Proceedings 14*, pages 231–240. Springer, 2020.

[4] Charitos Charitou, Simo Dragicevic, and Artur d'Avila Garcez. Synthetic data generation for fraud detection using gans. *arXiv preprint arXiv :2109.12546*, 2021.

[5] Nexialog Consulting. Méthodes de rééquilibrage des classes en classification supervisée. *working paper*, 2023.

[6] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11) :139–144, 2020.

[8] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2013.

[10] Belén Vega-Márquez, Cristina Rubio-Escudero, José C. Riquelme, and Isabel Nepomuceno-Chamorro. Creation of synthetic data with conditional generative adversarial networks. In Francisco Martínez Álvarez, Alicia Troncoso Lora, José António Sáez Muñoz, Héctor Quintián, and Emilio Corchado, editors, *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, pages 231–240, Cham, 2020. Springer International Publishing.

[11] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv :2112.09238*, 2021.

[12] Andrea Minieri. Synthetic data for privacy preservation - part 2. *clearbox.ai*, 2022.

[13] Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, 4 :679939, 2021.

[14] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 181(3) :663–688, 2018.

[15] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. Utility and privacy assessments of synthetic data for regression tasks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5763–5772. IEEE, 2019.

[16] SDmetrics. Quality metrics. *https ://-docs.sdv.dev/sdmetrics/metrics/*, 2023.

[17] Nexialog Consulting. Scoring par machine learning interprétable. *working paper*, 2023.

Nexialog Consulting est un cabinet de conseil spécialisé en Stratégie, Actuariat, Gestion des risques et Data qui dessert aujourd'hui les plus grands acteurs de la banque et de l'assurance Nous aidons nos clients à améliorer de manière significative et durable leurs performances et à atteindre leurs objectifs les plus importants.

Les besoins de nos clients et les réglementations européennes et mondiales étant en perpétuelle évolution, nous recherchons continuellement de nouvelles et meilleures façons de les servir. Pour ce faire, nous recrutons nos consultants dans les meilleures écoles d'ingénieur et de commerce et nous investissons des ressources de notre entreprise chaque année dans la recherche, l'apprentissage et le renforcement des compétences. Quelque soit le défi à relever, nous nous attachons à fournir des résultats pratiques et durables et à donner à nos clients les moyens de se développer.

Site web du cabinet : https://www.nexialog.com
Publications : https://www.nexialog.com/publications-nexialog/

Contacts

Ali BEHBAHANI
Associé, Fondateur
Tél : + 33 (0) 1 44 73 86 78
Email : abehbahani@nexialog.com

Christelle BONDOUX
Associée, Direction commerciale & Recrutement
Tél : + 33 (0) 1 44 73 75 67
Email : cbondoux@nexialog.com

Vivien BRUNEL
Associé, Data & Innovation
Tél : + 33 (0) 6 71 23 38 97
Email : vbrunel@nexialog.com

Areski COUSIN
Directeur scientifique
Tél : + 33 (0) 7 88 03 51 87
Email : acousin@nexialog.com